assignment_2_STUDENT

November 5, 2023

1 Assignment 2 (90 marks)

1.1 *The adverse health effects of air pollution - are we making any progress?*

Credit: Flickr/E4C

[38]:
```
# Load relevant packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as sm
import warnings

warnings.filterwarnings("ignore")  # Suppress all warnings
```

Introduction

Business Context. Air pollution is a very serious issue that the global population is currently dealing with. The abundance of air pollutants is not only contributing to global warming, but it is also causing problematic health issues to the population. There have been numerous efforts to protect and improve air quality across most nations. However, it seems that we are making very little progress. One of the main causes of this is the fact that the majority of air pollutants are derived from the burning of fossil fuels such as coal. Big industries and several other economical and political factors have slowed the progress towards the use of renewable energy by promoting the use of fossil fuels. Nevertheless, if we educate the general population and create awareness of this issue, we will be able to overcome this problem in the future.

For this case, you have been hired as a data science consultant for an important environmental organization. In order to promote awareness of environmental and greenhouse gas issues, your client is interested in a study of plausible impacts of air contamination on the health of the global population. They have gathered some raw data provided by the World Health Organization, The Institute for Health Metrics and Evaluation and the World Bank Group. Your task is to conduct data analysis, search for potential information, and create visualizations that the client can use for their campaigns and grant applications.

Analytical Context. You are given a folder, named files with raw data. This data contains quite a large number of variables and it is in a fairly disorganized state. In addition, one of the datasets contains very poor documentation, segmented into several datasets. Your objective will be to:

Extract and clean the relevant data. You will have to manipulate several datasets to obtain useful information for the case.

Conduct Exploratory Data Analysis. You will have to create meaningful plots, formulate meaningful hypotheses and study the relationship between various indicators related to air pollution.

Additionally, the client has some broad questions they would like to answer: 1. Are we making any progress in reducing the amount of emitted pollutants across the globe? 2. Which are the critical regions where we should start environmental campaigns? 3. Are we making any progress in the prevention of deaths related to air pollution? 4. Which demographic characteristics seem to correlate with the number of health-related issues derived from air pollution?

Extracting and cleaning relevant data

Let's take a look at the data provided by the client in the files folder. There, we see another folder named WDI_csv with several CSV files corresponding to the World Bank's primary World Development Indicators. The client stated that this data may contain some useful information relevant to our study, but they have not told us anything aside from that. Thus, we are on our own in finding and extracting the relevant data for our study. This we will do next.

Let's take a peek at the file WDIData.csv:

[39]:
```
WDI_data = pd.read_csv("./files/WDI_csv/WDIData.csv")
print(WDI_data.columns)
print(WDI_data.info())
WDI_data.head()
```

```
Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
       '1960', '1961', '1962', '1963', '1964', '1965', '1966', '1967', '1968',
       '1969', '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977',
       '1978', '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986',
       '1987', '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995',
       '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004',
       '2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
       '2014', '2015', '2016', '2017', '2018', '2019', 'Unnamed: 64'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 377256 entries, 0 to 377255
Data columns (total 65 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Country Name    377256 non-null  object
 1   Country Code    377256 non-null  object
 2   Indicator Name  377256 non-null  object
 3   Indicator Code  377256 non-null  object
 4   1960            37395 non-null   float64
 5   1961            41211 non-null   float64
 6   1962            43413 non-null   float64
 7   1963            43324 non-null   float64
 8   1964            43861 non-null   float64
```

9	1965	46306 non-null	float64
10	1966	46087 non-null	float64
11	1967	47840 non-null	float64
12	1968	47422 non-null	float64
13	1969	49112 non-null	float64
14	1970	69736 non-null	float64
15	1971	76073 non-null	float64
16	1972	78854 non-null	float64
17	1973	78402 non-null	float64
18	1974	79804 non-null	float64
19	1975	83728 non-null	float64
20	1976	85833 non-null	float64
21	1977	89303 non-null	float64
22	1978	88911 non-null	float64
23	1979	89707 non-null	float64
24	1980	94479 non-null	float64
25	1981	96363 non-null	float64
26	1982	97575 non-null	float64
27	1983	97385 non-null	float64
28	1984	98228 non-null	float64
29	1985	99450 non-null	float64
30	1986	100294 non-null	float64
31	1987	101654 non-null	float64
32	1988	101307 non-null	float64
33	1989	103060 non-null	float64
34	1990	126117 non-null	float64
35	1991	131212 non-null	float64
36	1992	135229 non-null	float64
37	1993	136645 non-null	float64
38	1994	138646 non-null	float64
39	1995	146560 non-null	float64
40	1996	146450 non-null	float64
41	1997	147530 non-null	float64
42	1998	149527 non-null	float64
43	1999	154659 non-null	float64
44	2000	179600 non-null	float64
45	2001	169874 non-null	float64
46	2002	174693 non-null	float64
47	2003	175686 non-null	float64
48	2004	180936 non-null	float64
49	2005	194452 non-null	float64
50	2006	192699 non-null	float64
51	2007	196798 non-null	float64
52	2008	195843 non-null	float64
53	2009	196888 non-null	float64
54	2010	211863 non-null	float64
55	2011	203080 non-null	float64
56	2012	204810 non-null	float64

```
 57  2013             200522 non-null  float64
 58  2014             206201 non-null  float64
 59  2015             201043 non-null  float64
 60  2016             197174 non-null  float64
 61  2017             176112 non-null  float64
 62  2018             126115 non-null  float64
 63  2019             21481 non-null   float64
 64  Unnamed: 64      0 non-null       float64
dtypes: float64(61), object(4)
memory usage: 187.1+ MB
None
```

[39]:
```
  Country Name Country Code  \
0   Arab World          ARB
1   Arab World          ARB
2   Arab World          ARB
3   Arab World          ARB
4   Arab World          ARB

                                      Indicator Name     Indicator Code  1960  \
0  2005 PPP conversion factor, GDP (LCU per inter…      PA.NUS.PPP.05   NaN
1  2005 PPP conversion factor, private consumptio…  PA.NUS.PRVT.PP.05   NaN
2  Access to clean fuels and technologies for coo…     EG.CFT.ACCS.ZS   NaN
3            Access to electricity (% of population)     EG.ELC.ACCS.ZS   NaN
4  Access to electricity, rural (% of rural popul…  EG.ELC.ACCS.RU.ZS   NaN

   1961  1962  1963  1964  1965  …       2011       2012       2013  \
0   NaN   NaN   NaN   NaN   NaN  …        NaN        NaN        NaN
1   NaN   NaN   NaN   NaN   NaN  …        NaN        NaN        NaN
2   NaN   NaN   NaN   NaN   NaN  …  82.783289  83.120303  83.533457
3   NaN   NaN   NaN   NaN   NaN  …  86.428272  87.070576  88.176836
4   NaN   NaN   NaN   NaN   NaN  …  73.942103  75.244104  77.162305

        2014       2015       2016       2017  2018  2019  Unnamed: 64
0        NaN        NaN        NaN        NaN   NaN   NaN          NaN
1        NaN        NaN        NaN        NaN   NaN   NaN          NaN
2  83.897596  84.171599  84.510171        NaN   NaN   NaN          NaN
3  87.342739  89.130121  89.678685  90.273687   NaN   NaN          NaN
4  75.538976  78.741152  79.665635  80.749293   NaN   NaN          NaN

[5 rows x 65 columns]
```

The data seems to have a large number of indicators dating from 1960. There are also columns containing country names and codes. Notice that the first couple of rows say Arab World, which may indicate that the data contains broad regional data as well. We notice also that there are at least 100,000 entries with NaN values for each year column.

Since we are interested in environmental indicators, we must get rid of any rows not relevant to our study. However, the number of indicators seems to be quite large and a manual inspection seems impossible. Let's load the file WDISeries.csv which seems to contain more information about the indicators:

[40]:
```
WDI_ids = pd.read_csv("./files/WDI_csv/WDISeries.csv")
print(WDI_ids.columns)
WDI_ids.head()
```

```
Index(['Series Code', 'Topic', 'Indicator Name', 'Short definition',
       'Long definition', 'Unit of measure', 'Periodicity', 'Base Period',
       'Other notes', 'Aggregation method', 'Limitations and exceptions',
       'Notes from original source', 'General comments', 'Source',
       'Statistical concept and methodology', 'Development relevance',
       'Related source links', 'Other web links', 'Related indicators',
       'License Type', 'Unnamed: 20'],
      dtype='object')
```

[40]:
```
         Series Code                                 Topic  \
0     AG.AGR.TRAC.NO  Environment: Agricultural production
1  AG.CON.FERT.PT.ZS  Environment: Agricultural production
2     AG.CON.FERT.ZS  Environment: Agricultural production
3     AG.LND.AGRI.K2                 Environment: Land use
4     AG.LND.AGRI.ZS                 Environment: Land use

                                      Indicator Name Short definition  \
0                   Agricultural machinery, tractors              NaN
1  Fertilizer consumption (% of fertilizer produc…              NaN
2  Fertilizer consumption (kilograms per hectare …              NaN
3                         Agricultural land (sq. km)              NaN
4                   Agricultural land (% of land area)              NaN

                                     Long definition Unit of measure  \
0  Agricultural machinery refers to the number of…             NaN
1  Fertilizer consumption measures the quantity o…             NaN
2  Fertilizer consumption measures the quantity o…             NaN
3  Agricultural land refers to the share of land …             NaN
4  Agricultural land refers to the share of land …             NaN

  Periodicity Base Period Other notes Aggregation method  …  \
0      Annual         NaN         NaN                Sum  …
1      Annual         NaN         NaN   Weighted average  …
2      Annual         NaN         NaN   Weighted average  …
3      Annual         NaN         NaN                Sum  …
4      Annual         NaN         NaN   Weighted average  …

  Notes from original source General comments  \
0                        NaN              NaN
1                        NaN              NaN
```

```
2                        NaN              NaN
3                        NaN              NaN
4                        NaN              NaN

                                              Source  \
0  Food and Agriculture Organization, electronic …
1  Food and Agriculture Organization, electronic …
2  Food and Agriculture Organization, electronic …
3  Food and Agriculture Organization, electronic …
4  Food and Agriculture Organization, electronic …

                 Statistical concept and methodology  \
0  A tractor provides the power and traction to m…
1  Fertilizer consumption measures the quantity o…
2  Fertilizer consumption measures the quantity o…
3  Agricultural land constitutes only a part of a…
4  Agriculture is still a major sector in many ec…

                               Development relevance Related source links  \
0  Agricultural land covers more than one-third o…                  NaN
1  Factors such as the green revolution, has led …                  NaN
2  Factors such as the green revolution, has led …                  NaN
3  Agricultural land covers more than one-third o…                  NaN
4  Agricultural land covers more than one-third o…                  NaN

  Other web links  Related indicators License Type Unnamed: 20
0             NaN                 NaN    CC BY-4.0         NaN
1             NaN                 NaN    CC BY-4.0         NaN
2             NaN                 NaN    CC BY-4.0         NaN
3             NaN                 NaN    CC BY-4.0         NaN
4             NaN                 NaN    CC BY-4.0         NaN

[5 rows x 21 columns]
```

Bingo! The WDI_ids DataFrame contains a column named Topic. Moreover, it seems that Environment is listed as a key topic in the column.

Exercise 1 (4 marks):

Extract all the rows that have the topic key Environment in WDI_ids. Add to the resulting DataFrame a new column named Subtopic which contains the corresponding subtopic of the indicator. For example, the subtopic of Environment: Agricultural production is Agricultural production. Which subtopics do you think are of interest to us?

Hint: Remember that you can apply string methods to Series using the str() method of pandas.

Answer.

[111]:
```python
# Extract rows with the topic "Environment: ***"
environment_data = WDI_ids[WDI_ids['Topic'].str.contains("Environment:")]

#create subtopic column
#(.+) is jused to capture captures any sequence of characters that exist after␣
↪the key topic "Environment: ".
#. matches any character, and * matches 0 to infinite occurrences of the␣
↪preceding character.
environment_data['Subtopic'] = environment_data['Topic'].str.
↪extract("Environment: (.*)")
# Display the subtopics
unique_subtopics = environment_data['Subtopic'].unique()
print("Subtopics:")
print(unique_subtopics)
environment_data.head(5)
```

```
Subtopics:
['Agricultural production' 'Land use' 'Energy production & use'
 'Emissions' 'Biodiversity & protected areas' 'Density & urbanization'
 'Freshwater' 'Natural resources contribution to GDP']
```

[111]:
```
        Series Code                                    Topic  \
0     AG.AGR.TRAC.NO  Environment: Agricultural production
1  AG.CON.FERT.PT.ZS  Environment: Agricultural production
2     AG.CON.FERT.ZS  Environment: Agricultural production
3     AG.LND.AGRI.K2                Environment: Land use
4     AG.LND.AGRI.ZS                Environment: Land use

                                      Indicator Name Short definition  \
0                  Agricultural machinery, tractors              NaN
1  Fertilizer consumption (% of fertilizer produc…              NaN
2  Fertilizer consumption (kilograms per hectare …              NaN
3                        Agricultural land (sq. km)              NaN
4                Agricultural land (% of land area)              NaN

                                     Long definition Unit of measure  \
0  Agricultural machinery refers to the number of…             NaN
1  Fertilizer consumption measures the quantity o…             NaN
2  Fertilizer consumption measures the quantity o…             NaN
3  Agricultural land refers to the share of land …             NaN
4  Agricultural land refers to the share of land …             NaN

  Periodicity Base Period Other notes Aggregation method  …  \
0      Annual         NaN         NaN                Sum  …
1      Annual         NaN         NaN   Weighted average  …
2      Annual         NaN         NaN   Weighted average  …
3      Annual         NaN         NaN                Sum  …
```

```
4      Annual         NaN         NaN   Weighted average  …

  General comments                                              Source  \
0              NaN  Food and Agriculture Organization, electronic …
1              NaN  Food and Agriculture Organization, electronic …
2              NaN  Food and Agriculture Organization, electronic …
3              NaN  Food and Agriculture Organization, electronic …
4              NaN  Food and Agriculture Organization, electronic …

                Statistical concept and methodology  \
0  A tractor provides the power and traction to m…
1  Fertilizer consumption measures the quantity o…
2  Fertilizer consumption measures the quantity o…
3  Agricultural land constitutes only a part of a…
4  Agriculture is still a major sector in many ec…

                              Development relevance Related source links  \
0  Agricultural land covers more than one-third o…                  NaN
1  Factors such as the green revolution, has led …                  NaN
2  Factors such as the green revolution, has led …                  NaN
3  Agricultural land covers more than one-third o…                  NaN
4  Agricultural land covers more than one-third o…                  NaN

  Other web links Related indicators  License Type Unnamed: 20  \
0             NaN                NaN     CC BY-4.0         NaN
1             NaN                NaN     CC BY-4.0         NaN
2             NaN                NaN     CC BY-4.0         NaN
3             NaN                NaN     CC BY-4.0         NaN
4             NaN                NaN     CC BY-4.0         NaN

                 Subtopic
0  Agricultural production
1  Agricultural production
2  Agricultural production
3                 Land use
4                 Land use

[5 rows x 22 columns]
```

Which subtopics do you think are of interest to us? FIX:

Based on the provided context and the client's objectives, the subtopics of interest should be related to air pollution and its health impacts. These subtopics could include Emissions, since it helps asses the progress in pollution reduction and energy production & use, since this subtopic might be relevant for exploring the impact of different energy sources on air pollution and health outcomes. I also think we can look at Natural resources contribution to GDP, to determine if there is a high use of fossil fuels.

Exercise 2 (4 marks):

Use the results of Exercise 1 to create a new DataFrame with the history of all emissions indicators for countries and major regions. Call this new DataFrame Emissions_df. How many emissions indicators are in the study?

[42]:
```python
# Filter data for the "Emissions" subtopic
WDIData = pd.read_csv("./files/WDI_csv/WDIData.csv")
emissions_indicators = environment_data[environment_data['Subtopic'].str.
 ↪contains('emissions',case=False)]

# Filter the temp_df DataFrame to include only rows that match the␣
 ↪indicator_name variable (indicator names with subtopic of Emissions)
Emissions_df = WDIData[WDIData['Indicator Code'].
 ↪isin(emissions_indicators['Series Code'])]

# Check how many emissions indicators are there?
num_emissions_indicators = len(Emissions_df['Indicator Code'].unique())

print("Number of emissions indicators in the study:", num_emissions_indicators)
Emissions_df
```

```
Number of emissions indicators in the study: 42
```

[42]:
```
       Country Name Country Code  \
64       Arab World          ARB
65       Arab World          ARB
66       Arab World          ARB
67       Arab World          ARB
191      Arab World          ARB
…               …            …
376814     Zimbabwe          ZWE
376815     Zimbabwe          ZWE
377064     Zimbabwe          ZWE
377160     Zimbabwe          ZWE
377161     Zimbabwe          ZWE

                                            Indicator Name  \
64            Agricultural methane emissions (% of total)
65      Agricultural methane emissions (thousand metri…
66      Agricultural nitrous oxide emissions (% of total)
67      Agricultural nitrous oxide emissions (thousand…
191                CO2 emissions (kg per 2010 US$ of GDP)
…                                                     …
376814  PM2.5 pollution, population exposed to levels …
376815  PM2.5 pollution, population exposed to levels …
```

```
377064  SF6 gas emissions (thousand metric tons of CO2…
377160  Total greenhouse gas emissions (% change from …
377161  Total greenhouse gas emissions (kt of CO2 equi…

            Indicator Code  1960  1961  1962  1963  1964  1965  …  \
64         EN.ATM.METH.AG.ZS   NaN   NaN   NaN   NaN   NaN   NaN  …
65      EN.ATM.METH.AG.KT.CE   NaN   NaN   NaN   NaN   NaN   NaN  …
66         EN.ATM.NOXE.AG.ZS   NaN   NaN   NaN   NaN   NaN   NaN  …
67      EN.ATM.NOXE.AG.KT.CE   NaN   NaN   NaN   NaN   NaN   NaN  …
191        EN.ATM.CO2E.KD.GD   NaN   NaN   NaN   NaN   NaN   NaN  …
…                        …   …   …   …   …   …   …   …
376814  EN.ATM.PM25.MC.T2.ZS   NaN   NaN   NaN   NaN   NaN   NaN  …
376815  EN.ATM.PM25.MC.T3.ZS   NaN   NaN   NaN   NaN   NaN   NaN  …
377064     EN.ATM.SF6G.KT.CE   NaN   NaN   NaN   NaN   NaN   NaN  …
377160        EN.ATM.GHGT.ZG   NaN   NaN   NaN   NaN   NaN   NaN  …
377161     EN.ATM.GHGT.KT.CE   NaN   NaN   NaN   NaN   NaN   NaN  …

                2011          2012        2013        2014        2015  \
64               NaN           NaN         NaN         NaN         NaN
65               NaN           NaN         NaN         NaN         NaN
66               NaN           NaN         NaN         NaN         NaN
67               NaN           NaN         NaN         NaN         NaN
191         0.757162      0.770413    0.737665    0.769023         NaN
…                …             …           …           …           …
376814     16.430216     22.112287   16.486892   18.625311    7.219464
376815    100.000000    100.000000  100.000000  100.000000  100.000000
377064           NaN           NaN         NaN         NaN         NaN
377160    103.876779    105.289436         NaN         NaN         NaN
377161  71561.952250  72057.803322         NaN         NaN         NaN

              2016       2017  2018  2019  Unnamed: 64
64             NaN        NaN   NaN   NaN          NaN
65             NaN        NaN   NaN   NaN          NaN
66             NaN        NaN   NaN   NaN          NaN
67             NaN        NaN   NaN   NaN          NaN
191            NaN        NaN   NaN   NaN          NaN
…              …         …   …   …           …
376814    8.708582    8.06692   NaN   NaN          NaN
376815  100.000000  100.00000   NaN   NaN          NaN
377064         NaN        NaN   NaN   NaN          NaN
377160         NaN        NaN   NaN   NaN          NaN
377161         NaN        NaN   NaN   NaN          NaN

[11088 rows x 65 columns]
```

Answer.

Exercise 3 (4 marks):

The DataFrame Emissions_df has one column per year of observation. Data in this form is usually referred to as data in wide format, as the number of columns is high. However, it might be easier to query and filter the data if we had a single column containing the year in which each indicator was calculated. This way, each observation will be represented by a single row. Use the pandas function melt() to reshape the Emissions_df data into long format. The resulting DataFrame should contain a pair of new columns named Year and Indicator Value:

Answer.

[113]:
```python
# reshape the Emissions_df DataFrame into long format
Emissions_df_long = Emissions_df.melt(
    id_vars=['Country Name', 'Country Code', 'Indicator Name', 'Indicator␣
↪Code'],
    var_name='Year',
    value_name='Indicator Value'
)

# Display the first few rows of the reshaped DataFrame
Emissions_df_long.head(15)
```

[113]:
```
   Country Name Country Code  \
0    Arab World          ARB
1    Arab World          ARB
2    Arab World          ARB
3    Arab World          ARB
4    Arab World          ARB
5    Arab World          ARB
6    Arab World          ARB
7    Arab World          ARB
8    Arab World          ARB
9    Arab World          ARB
10   Arab World          ARB
11   Arab World          ARB
12   Arab World          ARB
13   Arab World          ARB
14   Arab World          ARB

                                       Indicator Name        Indicator Code  \
0         Agricultural methane emissions (% of total)     EN.ATM.METH.AG.ZS
1   Agricultural methane emissions (thousand metri…  EN.ATM.METH.AG.KT.CE
2   Agricultural nitrous oxide emissions (% of total)     EN.ATM.NOXE.AG.ZS
3   Agricultural nitrous oxide emissions (thousand…  EN.ATM.NOXE.AG.KT.CE
4              CO2 emissions (kg per 2010 US$ of GDP)     EN.ATM.CO2E.KD.GD
5            CO2 emissions (kg per 2011 PPP $ of GDP)  EN.ATM.CO2E.PP.GD.KD
6                 CO2 emissions (kg per PPP $ of GDP)     EN.ATM.CO2E.PP.GD
7                                 CO2 emissions (kt)        EN.ATM.CO2E.KT
8              CO2 emissions (metric tons per capita)        EN.ATM.CO2E.PC
```

```
9   CO2 emissions from electricity and heat produc…       EN.CO2.ETOT.ZS
10  CO2 emissions from gaseous fuel consumption (%…     EN.ATM.CO2E.GF.ZS
11   CO2 emissions from gaseous fuel consumption (kt)     EN.ATM.CO2E.GF.KT
12  CO2 emissions from liquid fuel consumption (% …     EN.ATM.CO2E.LF.ZS
13    CO2 emissions from liquid fuel consumption (kt)     EN.ATM.CO2E.LF.KT
14  CO2 emissions from manufacturing industries an…       EN.CO2.MANF.ZS

    Year  Indicator Value
0   1960              NaN
1   1960              NaN
2   1960              NaN
3   1960              NaN
4   1960              NaN
5   1960              NaN
6   1960              NaN
7   1960     59535.396567
8   1960         0.645736
9   1960              NaN
10  1960         5.041292
11  1960              NaN
12  1960        84.851473
13  1960     50539.802737
14  1960              NaN
```

Exercise 4 (4 marks):

The column Indicator Value of the new Emissions_df contains a bunch of NaN values. Additionally, the Year column contains an Unnamed: 64 value. What procedure should we follow to clean these missing values in our DataFrame? Proceed with your suggested cleaning process.

Answer.

[114]:
```python
# Remove rows with missing values in the "Indicator Value" column
Emissions_df_cleaned = Emissions_df_long.dropna(subset=['Indicator Value'])

# Remove rows with "Year" equal to "Unnamed: 64"
Emissions_df_cleaned = Emissions_df_cleaned[Emissions_df_cleaned['Year'] !=␣
 ↪'Unnamed: 64']

Emissions_df_cleaned
```

[114]:
```
       Country Name Country Code  \
7        Arab World          ARB
8        Arab World          ARB
10       Arab World          ARB
12       Arab World          ARB
13       Arab World          ARB
```

```
…              …           …
643097     Zimbabwe         ZWE
643098     Zimbabwe         ZWE
643099     Zimbabwe         ZWE
643100     Zimbabwe         ZWE
652856        Sudan         SDN

                                              Indicator Name  \
7                                         CO2 emissions (kt)
8                   CO2 emissions (metric tons per capita)
10      CO2 emissions from gaseous fuel consumption (%…
12      CO2 emissions from liquid fuel consumption (% …
13        CO2 emissions from liquid fuel consumption (kt)
…                                                        …
643097  PM2.5 air pollution, population exposed to lev…
643098  PM2.5 pollution, population exposed to levels …
643099  PM2.5 pollution, population exposed to levels …
643100  PM2.5 pollution, population exposed to levels …
652856              CO2 emissions (metric tons per capita)

             Indicator Code  Year  Indicator Value
7              EN.ATM.CO2E.KT  1960     59535.396567
8              EN.ATM.CO2E.PC  1960         0.645736
10          EN.ATM.CO2E.GF.ZS  1960         5.041292
12          EN.ATM.CO2E.LF.ZS  1960        84.851473
13          EN.ATM.CO2E.LF.KT  1960     50539.802737
…                         …     …              …
643097      EN.ATM.PM25.MC.ZS  2017       100.000000
643098   EN.ATM.PM25.MC.T1.ZS  2017         0.000000
643099   EN.ATM.PM25.MC.T2.ZS  2017         8.066920
643100   EN.ATM.PM25.MC.T3.ZS  2017       100.000000
652856         EN.ATM.CO2E.PC  2018         0.000000

[325858 rows x 6 columns]
```

Exercise 5 (4 marks):

Split the Emissions_df into two DataFrames, one containing only countries and the other containing only regions. Name these Emissions_C_df and Emissions_R_df respectively.

Hint: You may want to inspect the file WDICountry.csv for this task. Region country codes may be found by looking at null values of the Region column in WDICountry.

Answer.

[116]:
```
WDICountry_df = pd.read_csv('./files/WDI_csv/WDICountry.csv')
```

```python
# Merge Emissions_df with WDICountry_df to add Region
Emissions_df_2 = Emissions_df_cleaned.merge(WDICountry_df[['Country Code',␣
 ↪'Region']], on='Country Code', how='left')

#Identify regions based on null values in the 'Region' column of WDICountry
Emissions_R_df = Emissions_df_2[Emissions_df_2['Region'].isnull()]
Emissions_R_df = Emissions_R_df.drop('Region', axis=1)

# Countries are the remaining rows in Emissions_df
Emissions_C_df = Emissions_df_2[Emissions_df_2['Region'].notna()]
Emissions_C_df = Emissions_C_df.drop('Region', axis=1)

Emissions_R_df
```

[116]:

```
                                  Country Name Country Code  \
0                                   Arab World          ARB
1                                   Arab World          ARB
2                                   Arab World          ARB
3                                   Arab World          ARB
4                                   Arab World          ARB
…                                            …            …
324881  Sub-Saharan Africa (IDA & IBRD countries)          TSS
324882                     Upper middle income          UMC
324883                     Upper middle income          UMC
324884                                   World          WLD
324885                                   World          WLD

                                          Indicator Name     Indicator Code  \
0                                    CO2 emissions (kt)     EN.ATM.CO2E.KT
1                  CO2 emissions (metric tons per capita)     EN.ATM.CO2E.PC
2       CO2 emissions from gaseous fuel consumption (%…  EN.ATM.CO2E.GF.ZS
3       CO2 emissions from liquid fuel consumption (% …  EN.ATM.CO2E.LF.ZS
4         CO2 emissions from liquid fuel consumption (kt)  EN.ATM.CO2E.LF.KT
…                                                     …                  …
324881  PM2.5 air pollution, population exposed to lev…  EN.ATM.PM25.MC.ZS
324882  PM2.5 air pollution, mean annual exposure (mic…  EN.ATM.PM25.MC.M3
324883  PM2.5 air pollution, population exposed to lev…  EN.ATM.PM25.MC.ZS
324884  PM2.5 air pollution, mean annual exposure (mic…  EN.ATM.PM25.MC.M3
324885  PM2.5 air pollution, population exposed to lev…  EN.ATM.PM25.MC.ZS

        Year  Indicator Value
0       1960     59535.396567
1       1960         0.645736
2       1960         5.041292
3       1960        84.851473
4       1960     50539.802737
…          …                …
```

```
324881  2017       100.000000
324882  2017        38.748285
324883  2017        96.065069
324884  2017        45.521859
324885  2017        91.295708

[62902 rows x 6 columns]
```

[118]:
```
Emissions_C_df
```

[118]:
```
       Country Name Country Code  \
450     Afghanistan          AFG
451     Afghanistan          AFG
452     Afghanistan          AFG
453     Afghanistan          AFG
454     Afghanistan          AFG
…             …           …
325853     Zimbabwe          ZWE
325854     Zimbabwe          ZWE
325855     Zimbabwe          ZWE
325856     Zimbabwe          ZWE
325857        Sudan          SDN

                                           Indicator Name  \
450                                     CO2 emissions (kt)
451                 CO2 emissions (metric tons per capita)
452     CO2 emissions from gaseous fuel consumption (%…
453      CO2 emissions from gaseous fuel consumption (kt)
454     CO2 emissions from liquid fuel consumption (% …
…                                                    …
325853  PM2.5 air pollution, population exposed to lev…
325854  PM2.5 pollution, population exposed to levels …
325855  PM2.5 pollution, population exposed to levels …
325856  PM2.5 pollution, population exposed to levels …
325857              CO2 emissions (metric tons per capita)

             Indicator Code  Year  Indicator Value
450            EN.ATM.CO2E.KT  1960       414.371000
451            EN.ATM.CO2E.PC  1960         0.046057
452         EN.ATM.CO2E.GF.ZS  1960         0.000000
453         EN.ATM.CO2E.GF.KT  1960         0.000000
454         EN.ATM.CO2E.LF.ZS  1960        65.486726
…                         …   …            …
325853      EN.ATM.PM25.MC.ZS  2017       100.000000
325854   EN.ATM.PM25.MC.T1.ZS  2017         0.000000
325855   EN.ATM.PM25.MC.T2.ZS  2017         8.066920
325856   EN.ATM.PM25.MC.T3.ZS  2017       100.000000
```

```
325857        EN.ATM.CO2E.PC  2018        0.000000

[262956 rows x 6 columns]
```

Finalizing the cleaning for our study

Our data has improved a lot by now. However, since the number of indicators is still quite large, let us focus our study on the following indicators for now:

Total greenhouse gas emissions (kt of CO2 equivalent), EN.ATM.GHGT.KT.CE: The total of greenhouse emissions includes CO2, Methane, Nitrous oxide, among other pollutant gases. Measured in kilotons.

CO2 emissions (kt), EN.ATM.CO2E.KT: Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.

Methane emissions (kt of CO2 equivalent), EN.ATM.METH.KT.CE: Methane emissions are those stemming from human activities such as agriculture and from industrial methane production.

Nitrous oxide emissions (kt of CO2 equivalent), EN.ATM.NOXE.KT.CE: Nitrous oxide emissions are emissions from agricultural biomass burning, industrial activities, and livestock management.

Other greenhouse gas emissions, HFC, PFC and SF6 (kt of CO2 equivalent), EN.ATM.GHGO.KT.CE: Other pollutant gases.

PM2.5 air pollution, mean annual exposure (micrograms per cubic meter), EN.ATM.PM25.MC.M3: Population-weighted exposure to ambient PM2.5 pollution is defined as the average level of exposure of a nation's population to concentrations of suspended particles measuring less than 2.5 microns in aerodynamic diameter, which are capable of penetrating deep into the respiratory tract and causing severe health damage. Exposure is calculated by weighting mean annual concentrations of PM2.5 by population in both urban and rural areas.

PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total), EN.ATM.PM25.MC.ZS: Percent of population exposed to ambient concentrations of PM2.5 that exceed the World Health Organization (WHO) guideline value.

Exercise 6 (5 marks) :

For each of the emissions DataFrames, extract the rows corresponding to the above indicators of interest. Replace the long names of the indicators by the short names Total, CO2, CH4, N2O, Other, PM2.5, and PM2.5_WHO. (This will be helpful later when we need to label plots of our data.)

Answer.

[119]:
```python
# Define a dictionary to map long indicator names to short names
indicator_mapping = {
    'EN.ATM.GHGT.KT.CE': 'Total',
    'EN.ATM.CO2E.KT': 'CO2',
    'EN.ATM.METH.KT.CE': 'CH4',
```

```
    'EN.ATM.NOXE.KT.CE': 'N2O',
    'EN.ATM.GHGO.KT.CE': 'Other',
    'EN.ATM.PM25.MC.M3': 'PM2.5',
    'EN.ATM.PM25.MC.ZS': 'PM2.5_WHO'
}
Emissions_R_filtered = Emissions_R_df.copy()
Emissions_C_filtered = Emissions_C_df.copy()
# Replace long indicator names with short names in Emissions_R_df
Emissions_R_filtered['Indicator Name'] = Emissions_R_df['Indicator Code'].
 ↪map(indicator_mapping)
# Replace long indicator names with short names in Emissions_C_df
Emissions_C_filtered['Indicator Name'] = Emissions_C_df['Indicator Code'].
 ↪map(indicator_mapping)
# Drop rows with NaN values in the 'Indicator Name' column
Emissions_R_filtered = Emissions_R_filtered.dropna(subset=['Indicator Name'])
Emissions_C_filtered = Emissions_C_filtered.dropna(subset=['Indicator Name'])
```

[120]:

```
Emissions_R_filtered
```

[120]:

```
                                   Country Name Country Code Indicator Name  \
0                                    Arab World          ARB            CO2
6                        Caribbean small states          CSS            CO2
12                 Central Europe and the Baltics          CEB            CO2
26                     Early-demographic dividend          EAR            CO2
40                          East Asia & Pacific          EAS            CO2
…                                             …            …              …
324881  Sub-Saharan Africa (IDA & IBRD countries)          TSS      PM2.5_WHO
324882                          Upper middle income          UMC          PM2.5
324883                          Upper middle income          UMC      PM2.5_WHO
324884                                        World          WLD          PM2.5
324885                                        World          WLD      PM2.5_WHO

           Indicator Code  Year  Indicator Value
0          EN.ATM.CO2E.KT  1960     5.953540e+04
6          EN.ATM.CO2E.KT  1960     5.878201e+03
12         EN.ATM.CO2E.KT  1960     4.665334e+05
26         EN.ATM.CO2E.KT  1960     5.821834e+05
40         EN.ATM.CO2E.KT  1960     1.210072e+06
…                     …    …              …
324881  EN.ATM.PM25.MC.ZS  2017     1.000000e+02
324882  EN.ATM.PM25.MC.M3  2017     3.874829e+01
324883  EN.ATM.PM25.MC.ZS  2017     9.606507e+01
324884  EN.ATM.PM25.MC.M3  2017     4.552186e+01
324885  EN.ATM.PM25.MC.ZS  2017     9.129571e+01

[11419 rows x 6 columns]
```

[121]:
```
Emissions_C_filtered
```

[121]:
```
                 Country Name Country Code Indicator Name     Indicator Code  \
450               Afghanistan          AFG            CO2     EN.ATM.CO2E.KT
458                   Albania          ALB            CO2     EN.ATM.CO2E.KT
467                   Algeria          DZA            CO2     EN.ATM.CO2E.KT
475                    Angola          AGO            CO2     EN.ATM.CO2E.KT
483       Antigua and Barbuda          ATG            CO2     EN.ATM.CO2E.KT
…                         …          …              …                …
325843            Yemen, Rep.          YEM      PM2.5_WHO  EN.ATM.PM25.MC.ZS
325847                 Zambia          ZMB          PM2.5  EN.ATM.PM25.MC.M3
325848                 Zambia          ZMB      PM2.5_WHO  EN.ATM.PM25.MC.ZS
325852               Zimbabwe          ZWE          PM2.5  EN.ATM.PM25.MC.M3
325853               Zimbabwe          ZWE      PM2.5_WHO  EN.ATM.PM25.MC.ZS

        Year  Indicator Value
450     1960       414.371000
458     1960      2024.184000
467     1960      6160.560000
475     1960       550.050000
483     1960        36.670000
…        …              …
325843  2017       100.000000
325847  2017        27.438035
325848  2017       100.000000
325852  2017        22.251671
325853  2017       100.000000

[48059 rows x 6 columns]
```

Where shall the client start environmental campaigns?

Now the DataFrames Emissions_C_df and Emissions_R_df seem to be in a good shape. Let's proceed to conduct some exploratory data analysis so that we can make recommendations to our client.

Exercise 7 (15 marks):

Let's first calculate some basic information about the main indicators across the globe.

7.1 (5 marks)

Compute some basic statistics of the amount of kt of emissions for each of the four main pollutants (CO2, CH4, N2O, Others) over the years. Use the Emissions_C_df data frame. What trends do you see?

[123]:
```
# Filter the Emissions_C_df DataFrame to select only the rows with the four␣
↪main pollutants
```

```
main_pollutants = ['CO2', 'CH4', 'N2O', 'Other']
filtered_df = Emissions_C_filtered[Emissions_C_filtered['Indicator Name'].
 ↪isin(main_pollutants)]

# Group the data by 'Indicator Name' (pollutants) and calculate basic␣
 ↪statistics for each pollutant
pollutant_stats = filtered_df.groupby('Indicator Name')['Indicator Value'].
 ↪describe()

# Display the basic statistics for the pollutants
print(pollutant_stats)
```

```
                 count           mean            std         min         25%  \
Indicator Name
CH4             8736.0   31900.185639  104985.622926       0.000  880.621250
CO2             9856.0  100481.131586  495094.173851     -80.674  557.384000
N2O             8779.0   13575.872976   41248.850927       0.000  291.106585
Other           7971.0   30824.989016  132149.566564 -326272.600    7.548464

                      50%           75%           max
Indicator Name
CH4             5457.5050  19325.339000  1.752290e+06
CO2             4275.7220  40085.810500  1.029193e+07
N2O             2499.2944   8913.466837  5.871664e+05
Other            843.2500  10754.864020  3.484920e+06
```

The first trend that I see is that all of these indicators have right-skewed distributions with high outliars. I alos noticed that the average levels of emissions are relatively high for all indicators, but especially for CO2 and CH4. I noticed that some of the minimum values are negative, this may mean that the dataset may contain anomalies which need to be investigated further. I also noticed that there is a pretty wide range in emissions for each indicator, suggesting that there are variations in emissions across countries and years (from std).

7.2 (3 marks)

What can you say about the distribution of emissions around the globe over the years? What information can you extract from the tails of these distributions over the years?

Answer.

[52]:
```
#ANSWERRRR
```

7.3 (7 marks)

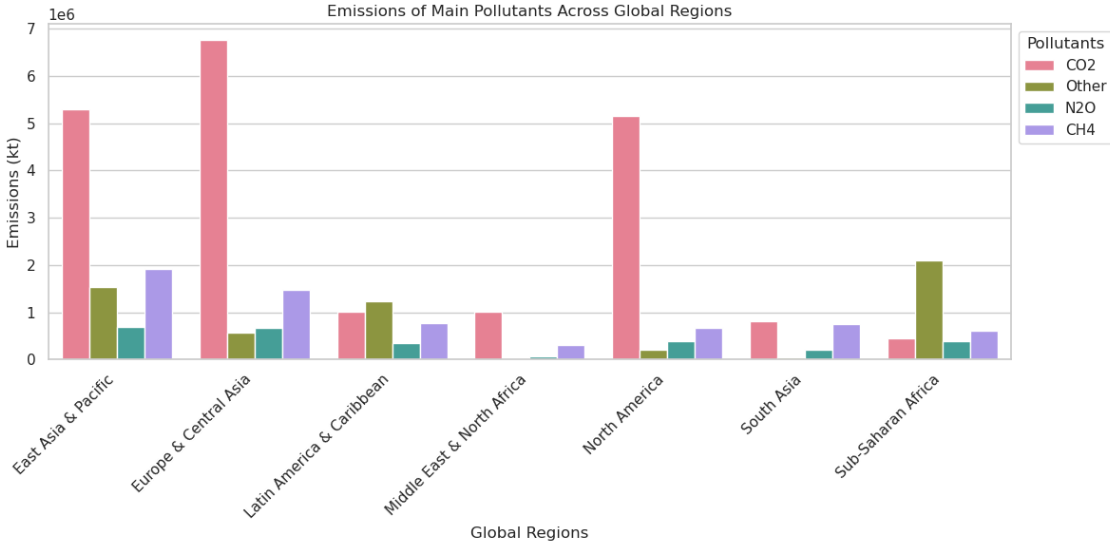
Compute a plot showing the behavior of each of the four main air pollutants for each of the main global regions in the Emissions_R_df data frame. The main regions are ‘Latin America &

Caribbean', 'South Asia', 'Sub-Saharan Africa', 'Europe & Central Asia', 'Middle East & North Africa', 'East Asia & Pacific' and 'North America'. What conclusions can you make?

Answer.

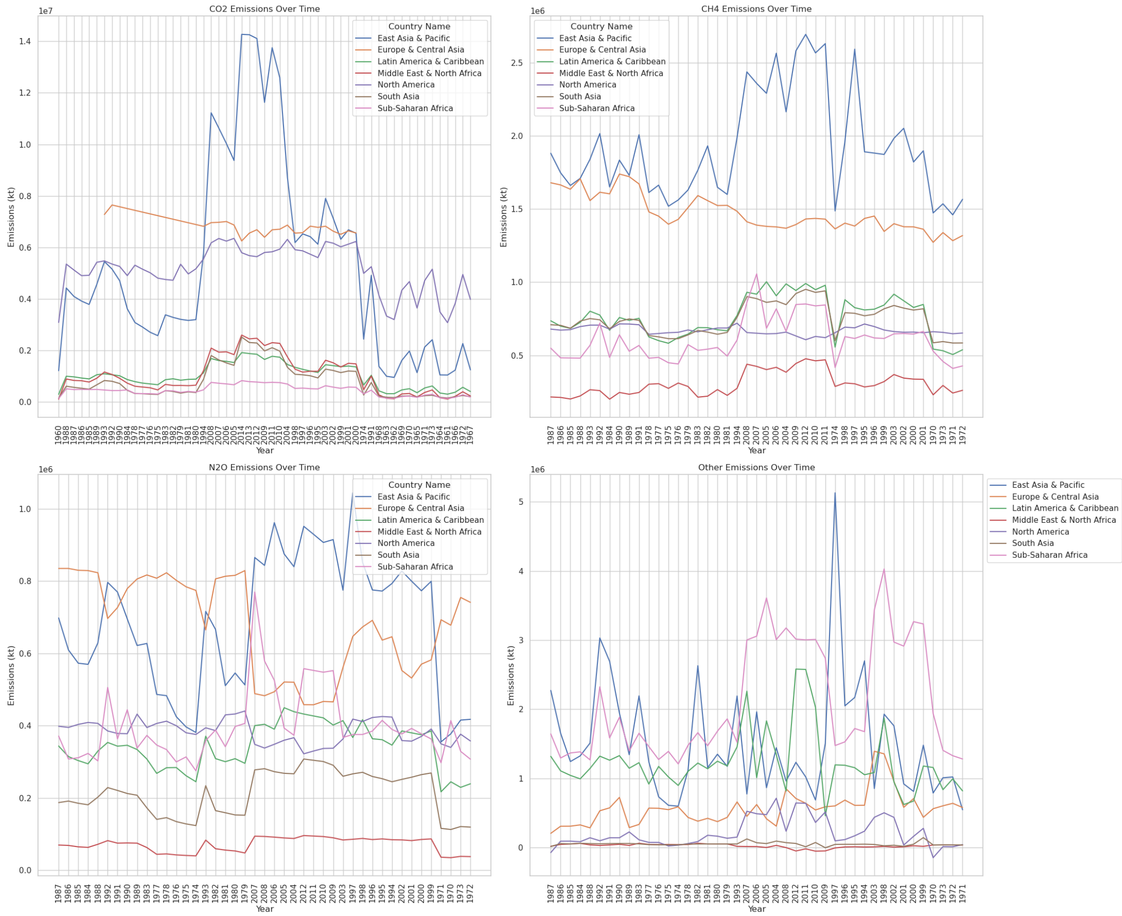
[125]:
```
# Filter the data for the main regions
main_regions = ['Latin America & Caribbean', 'South Asia', 'Sub-Saharan␣
↪Africa', 'Europe & Central Asia', 'Middle East & North Africa', 'East Asia &␣
↪Pacific', 'North America']
# Merge Emissions_R_df with WDICountry_df to add the 'Region' column
filtered_df = Emissions_R_filtered[Emissions_R_filtered['Country Name'].
↪isin(main_regions)]
filtered_df = filtered_df.dropna(subset=['Indicator Name'])
# Filter the DataFrame to include only rows with these pollutants
filtered_df = filtered_df[filtered_df['Indicator Name'].isin(main_pollutants)]
```

[126]:
```
plt.figure(figsize=(12, 6))
sns.set(style="whitegrid")

# Sort the data by 'Country Name' in alphabetical order
filtered_df = filtered_df.sort_values(by='Country Name')

sns.barplot(x='Country Name', y='Indicator Value', hue='Indicator Name',␣
↪data=filtered_df, palette='husl', ci=None)

plt.title('Emissions of Main Pollutants Across Global Regions')
plt.xlabel('Global Regions')
plt.ylabel('Emissions (kt)')

# Rotate x-axis labels for better readability
plt.xticks(rotation=45, ha='right')

plt.legend(title='Pollutants', bbox_to_anchor=(1, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



[129]:

```
# Create a plot for each pollutant to compare overtime
plt.figure(figsize=(22, 18))  # Larger figsize

for i, pollutant in enumerate(main_pollutants):
    plt.subplot(2, 2, i + 1)
    sns.lineplot(x='Year', y='Indicator Value', hue='Country Name',␣
 ↪data=filtered_df[filtered_df['Indicator Name'] == pollutant])
    plt.title(f'{pollutant} Emissions Over Time', fontsize=12)
    plt.xlabel('Year')
    plt.ylabel('Emissions (kt)')

    # Slant the titles
    plt.xticks(rotation=90)

# Add a legend outside the subplots
plt.legend(loc='upper left', bbox_to_anchor=(1, 1))

plt.tight_layout()
plt.show()
```



From these line plots, we can see that countries in East Asia and the Pacific are the worst dealing with pollutant emissions. This is because compared to all of the other regions, they have higher emissions overall for all 4 indicators. However, there are also some regions that are making improvements. For example, we can see that the region of Europe & Central Asia has been making imporvements over time for both C02 and CH4 emissions.

Exercise 8 (10 marks):

In Exercise 7 we discovered some interesting features of the distribution of the emissions over the years. Let us explore these features in more detail.

8.1 (5 marks)

Which are the top five countries that have been in the top 10 of CO2 emitters over the years? Have any of these countries made efforts to reduce the amount of CO2 emissions over the last 10 years?

Answer.

[56]:
```
# 1. Calculate the total CO2 emissions for each country all time
```

```
co2_emissions = Emissions_C_filtered[Emissions_C_filtered['Indicator Name'] ==␣
↪'CO2']
total_co2_emissions_by_country = co2_emissions.groupby('Country␣
↪Name')['Indicator Value'].sum().reset_index()

# 2. Sort the countries based on their total CO2 emissions
top_emitters = total_co2_emissions_by_country.sort_values(by='Indicator Value',␣
↪ascending=False)

# 3. Select the top five countries
top_5_emitters = top_emitters.head(5)

# Display the top five countries that have been in the top 10 of CO2 emitters␣
↪over the years
top_5_emitters
```

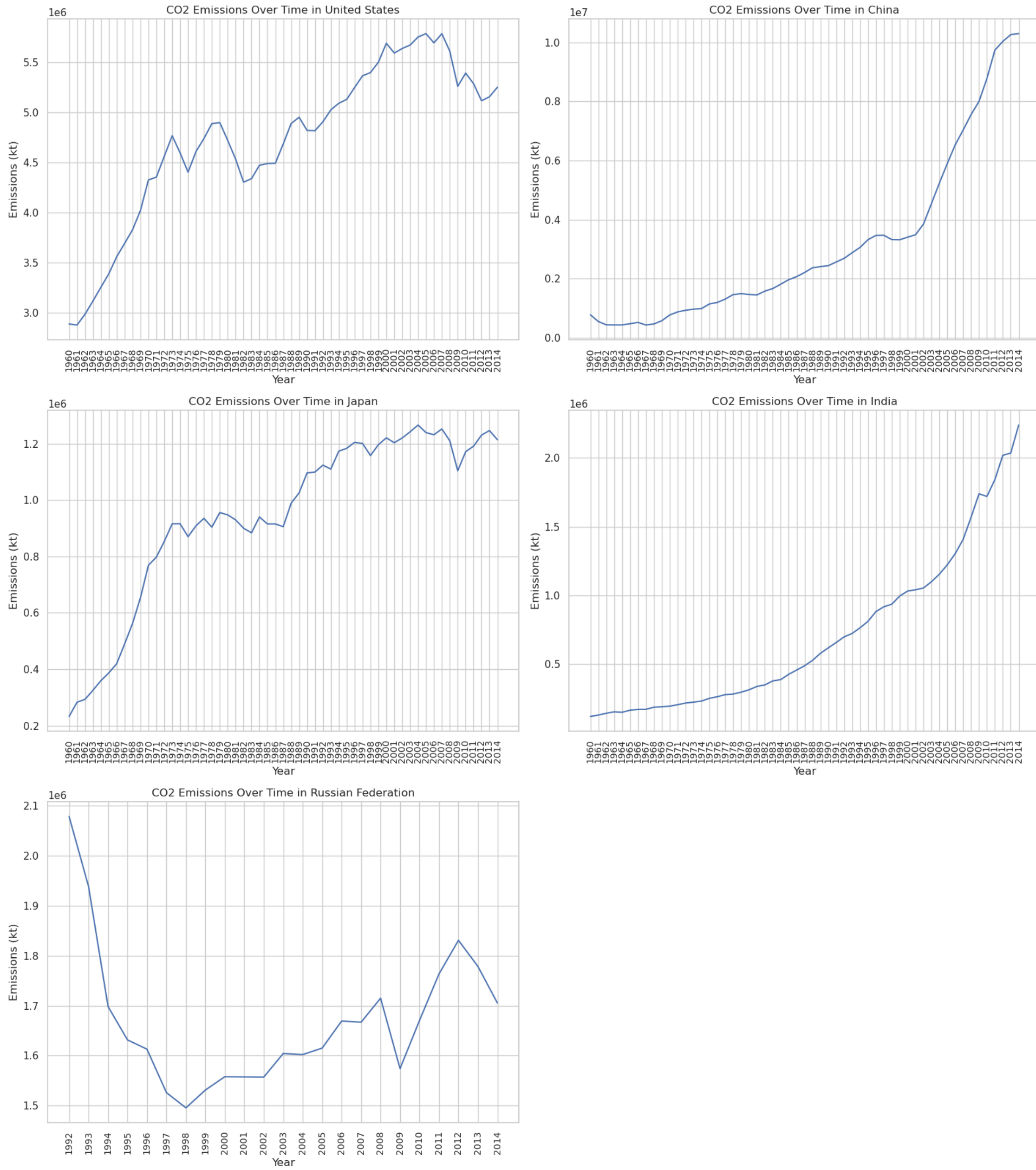
[56]:

```
          Country Name  Indicator Value
196      United States     2.597893e+08
39               China     1.704215e+08
94               Japan     5.197259e+07
86               India     3.876964e+07
153 Russian Federation     3.838037e+07
```

[130]:

```
# List of countries for which you want to create plots
countries = ['United States', 'China', 'Japan', 'India', 'Russian Federation']

# Filter the data for CO2 emissions for the selected countries and create␣
↪separate plots
plt.figure(figsize=(16, 18))  # Adjust the figure size as needed

for i, country in enumerate(countries):
    plt.subplot(3, 2, i + 1)  # 3 rows and 2 columns

    country_co2_emissions =␣
↪Emissions_C_filtered[(Emissions_C_filtered['Indicator Name'] == 'CO2') &␣
↪(Emissions_C_filtered['Country Name'] == country)]

    sns.lineplot(x='Year', y='Indicator Value', data=country_co2_emissions)
    plt.title(f'CO2 Emissions Over Time in {country}', fontsize=12)
    plt.xlabel('Year')
    plt.ylabel('Emissions (kt)')

    # Rotate the x-axis labels for better readability and adjust font size
    plt.xticks(rotation=90, fontsize=10)

plt.tight_layout()
```

```
plt.show()
```



Based on the plots above I can determine if any of the top 5 countries have made improvements over the last 10 years. For USA, there has been a drop in emissions since 2007 so I'd think that they have made many efforts to lower their emissions. Russia is on the verge of improving however they need to show more improvent to compensate for their high usage in 2009-2012. As for India and China, there is no signs of decrease in emissions whatsoever in the last 10 years which is very dissapointing. Japan is not increasing its emissions by as much as India and China, however it still ius not trying to decrease its emissions either in last 10 years.

8.2 (5 marks)

Are these five countries carrying out the burden of most of the emissions emitted over the years globally? Can we say that the rest of the world is making some effort to control their polluted gasses emissions over the years?

Answer.

[58]:
```
# Sort the countries based on their total CO2 emissions
all_emitters = total_co2_emissions_by_country.sort_values(by='Indicator Value',␣
↪ascending=False)

total_global_co2_emissions = all_emitters['Indicator Value'].sum()


top_5_percentage = (top_5_emitters['Indicator Value'].sum() /␣
↪total_global_co2_emissions) * 100

rest_of_world_percentage = 100 - top_5_percentage

# Display the results
print(f'Total Global CO2 Emissions: {total_global_co2_emissions} kt')
print(f'Total CO2 Emissions from the Top 5 Emitters: {top_5_emitters["Indicator␣
↪Value"].sum()} kt')
print(f'Percentage of Global Emissions by Top 5 Emitters: {top_5_percentage:.
↪2f}%')
print(f'Percentage of Global Emissions by the Rest of the World:␣
↪{rest_of_world_percentage:.2f}%')
```

```
Total Global CO2 Emissions: 990342032.91 kt
Total CO2 Emissions from the Top 5 Emitters: 559333424.871 kt
Percentage of Global Emissions by Top 5 Emitters: 56.48%
Percentage of Global Emissions by the Rest of the World: 43.52%
```

Based on the info above, I do think that the top 5 countries are carrying out the burden of most of the emissions emitted over the years globally, since their percentage towards gloabkl emissions is 56.48%! That's a very large proportion for only 5 countries to take up. If these 5 countries try to seriosuly reduce their emissions, the global emissions will go down by a lot as a direct result.

The health impacts of air pollution

Exercise 9 (10 marks):

One of the main contributions of poor health from air pollution is particulate matter. In particular, very small particles (those with a size less than 2.5 micrometres (μm)) can enter and affect the respiratory system. The PM2.5 indicator measures the average level of exposure of a nation's population to concentrations of these small particles. The PM2.5_WHO measures the percentage of the population who are exposed to ambient concentrations of these particles that exceed some thresholds set by the World Health Organization (WHO). In particular, countries with a higher PM2.5_WHO indicator are more likely to suffer from bad health conditions.

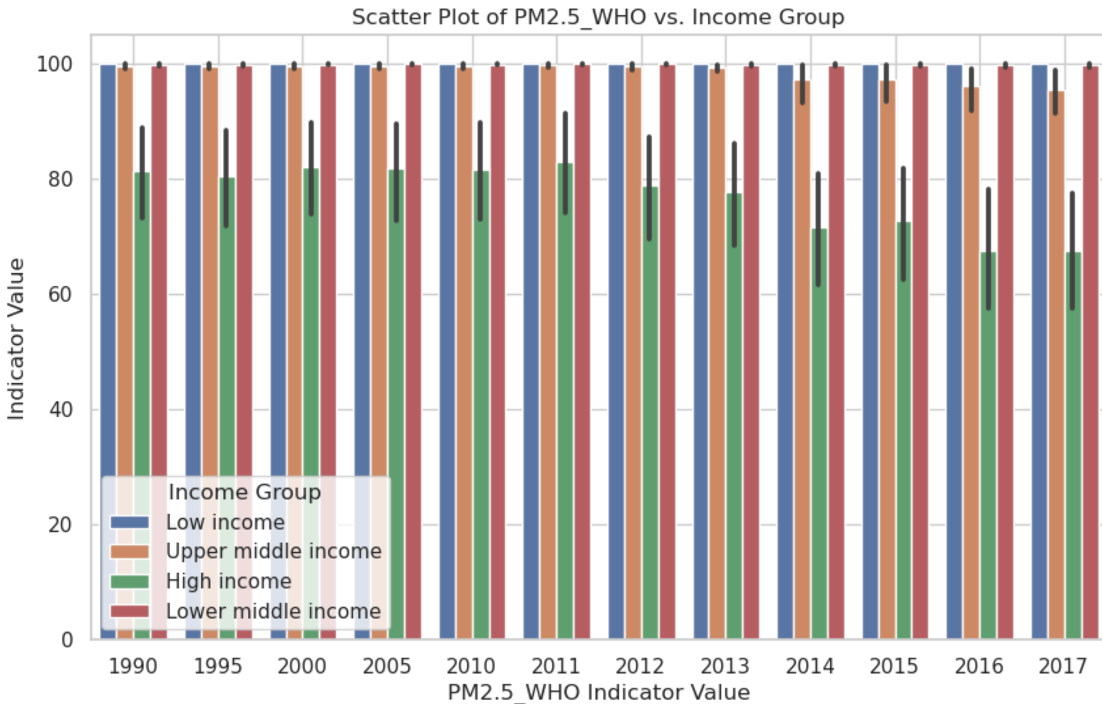
9.1 (7 marks)

The client would like to know if there is any relationship between the PM2.5_WHO indicator and the level of income of the general population, as well as how this changes over time. What plot(s) might be helpful to solve the client's question? What conclusion can you draw from your plot(s) to answer their question?

Hint: The DataFrame WDI_countries contains a column named Income Group.

Answer.

[70]:
```python
WDI_countries = pd.read_csv("./files/WDI_csv/WDICountry.csv")

merged_emissions = pd.concat([Emissions_R_filtered, Emissions_C_filtered],␣
 ↪axis=0)

# Filter the data for the PM2.5_WHO indicator
pm25_data = merged_emissions[merged_emissions['Indicator Name'] == 'PM2.5_WHO']

# Merge the PM2.5_WHO data with the income group data
pm25_with_income = pm25_data.merge(WDI_countries[['Country Code', 'Income␣
 ↪Group']], on='Country Code')

# Create a scatter plot
plt.figure(figsize=(10, 6))
sns.barplot(x='Year', y='Indicator Value', hue='Income Group',␣
 ↪data=pm25_with_income)

# Add labels and a title
plt.xlabel('PM2.5_WHO Indicator Value')
plt.ylabel('Indicator Value')
plt.title('Scatter Plot of PM2.5_WHO vs. Income Group')

# Show the plot
plt.grid(True)
plt.show()

pm25_with_income.describe()
```



[70]:

	Indicator Value
count	2880.000000
mean	92.952570
std	21.202702
min	0.000000
25%	99.384072
50%	100.000000
75%	100.000000
max	100.000000

I think that a grouped bar plot is a helpful choice for visualizing the relationship between the PM2.5_WHO indicator and the level of income (Income Group) of the general population over time. It is suitable for answering the client's question about the relationship between these variables and how it changes over time, since we can use hue to show the indicator values for each income group, for all years that are recorded. Grouped bar plots allow you to compare and contrast multiple categories (income groups) simultaneously for each year, allows you to observe change/progress over time, it makes it clear how different income groups are affected by particulate matter pollution,

What conclusion can you draw from your plot(s) to answer their question?

The main and most obvious conclusion that I can draw from this grouped bar plot, is that the population that is classified into the low income bracket, is almost gauranteed to be exposed to ambient concentrations of fine particulate matter that exceed the thresholds set by WHO. Specifically, throughout all the years shown in the graph, 100% of the low income population is exposed to PM2.5 at a concentration level that WHO consideres to be very dangerous. I also see that the lower-middle income group is very close to having 100% of the population being exposed to PM2.5 at a level such that it can cause you to suffer from bad health conditions. Lower-middle income group and lower income group have a very similar trend throguhotu the years (high exposure percentage and stay consistent). As for the upper-middle income group, even though the percentage above the hazard level specified by WHO is still very high, we can see that progress is made slowly over the years. Specifically, you can see that during the years 2014-2017, the perctange of the population being exposed to a hazardous amoutn of PM2.5, is decreasing while still greater than 90%. Laslty, the high income population is much more fortunate than the rest since they have less peole experiencing the hazardous level of PM2.5 that was specified by WHO. This income group experiences slight increases and slight decreases in PM2.5 perctange over the years. However, years 2011-2014 show the most cosistent time period of decrease. The most recent years recorded also have the lowest percentages, which is a good sign of making progress in the right direction.

9.2 (3 marks)

What do you think are the causes behind the results in Exercise 9.1?

Answer.

I think that there are many underlying causes behind the results in 9.1. First off, is the affect of income disparities. I think that lower income groups are more likely to experience high PM2.5 exposure due to limited access to clean energy. Whereas on the flip side, higher income groups benefit from economic development, cleaner technologies, and improved infrastructure, leading to reduced pollution exposure. Also, low-income countries may have weaker environmental regulations, which leads to higher pollution. Access to healthcare is also a really big factor, since high income groups have better access, mitigating the health effects of pollution. Global and regional factors such as the climate and atmospheric conditions can also impact the air quality, while affecting different income groups.

Exercise 10 (30 marks):

Finally, our client is interested in investigating the impacts and relationships between high levels of exposure to particle matter and the health of the population. Coming up with additional data for this task may be infeasible for the client, thus they have asked us to search for relevant health data in the WDIdata.csv file and work with that.

10.1 (4 marks)

Which indicators present in the file WDISeries.csv file might be useful to solve the client's question? Explain.

Note: Naming one or two indicators is more than enough for this question.

Answer.

[60]:
```
WDI_ids_Mortality = WDI_ids[WDI_ids['Topic'].str.contains("Mortality")]
names = WDI_ids_Mortality['Indicator Name']
#984 #1013
names
```

```
[60]: 930                          Number of deaths ages 5-14 years
932                                    Number of infant deaths
934                                Number of under-five deaths
936                                  Number of neonatal deaths
937      Probability of dying at age 5-14 years (per 1,…
940        Mortality rate, under-5 (per 1,000 live births)
941      Mortality rate, under-5, female (per 1,000 liv…
942      Mortality rate, under-5, male (per 1,000 live …
943      Mortality from CVD, cancer, diabetes or CRD be…
944      Mortality from CVD, cancer, diabetes or CRD be…
945      Mortality from CVD, cancer, diabetes or CRD be…
946       Mortality rate, neonatal (per 1,000 live births)
984      Mortality rate attributed to household and amb…
985      Mortality rate attributed to household and amb…
986      Mortality rate attributed to household and amb…
1013     Mortality rate attributed to unintentional poi…
1014     Mortality rate attributed to unintentional poi…
1015     Mortality rate attributed to unintentional poi…
1022     Suicide mortality rate, female (per 100,000 fe…
1023     Suicide mortality rate, male (per 100,000 male…
1024       Suicide mortality rate (per 100,000 population)
1025     Mortality caused by road traffic injury (per 1…
1026     Mortality rate attributed to unsafe water, uns…
1231     Mortality rate, adult, female (per 1,000 femal…
1232     Mortality rate, adult, male (per 1,000 male ad…
1237     Mortality rate, infant, female (per 1,000 live…
1238        Mortality rate, infant (per 1,000 live births)
1239     Mortality rate, infant, male (per 1,000 live b…
1240              Life expectancy at birth, female (years)
1241               Life expectancy at birth, total (years)
1242                Life expectancy at birth, male (years)
1244              Survival to age 65, female (% of cohort)
1245                Survival to age 65, male (% of cohort)
Name: Indicator Name, dtype: object
```

Based on the WDI_ids_Mortality data frame, I think that the two indicators that could be helpful for solving the client's question, are "Mortality rate attributed to household and ambient air pollution, age-standardized (per 100,000 population)" and "Mortality rate attributed to unintentional poisoning (per 100,000 population)". This is because they relate to air pollution and unintential poisoning (potentially pullution poisoning).

10.2 (4 marks)

Use the indicators provided in Exercise 10.1 to give valuable information to the client.

Answer.

[135]:
```
indicator1 = WDI_ids_Mortality[WDI_ids_Mortality['Indicator Name'].str.
 ↪contains("Mortality rate attributed to unintentional poisoning ")]
indicator2 = WDI_ids_Mortality[WDI_ids_Mortality['Indicator Name'].str.
 ↪contains("Mortality rate attributed to household and ambient air pollution,␣
 ↪age-standardized ")]
indicator1
```

[135]:
```
        Series Code              Topic  \
1013  SH.STA.POIS.P5  Health: Mortality

                                         Indicator Name Short definition  \
1013  Mortality rate attributed to unintentional poi…              NaN

                                        Long definition Unit of measure  \
1013  Mortality rate attributed to unintentional poi…             NaN

     Periodicity Base Period Other notes Aggregation method  …  \
1013      Annual         NaN         NaN   Weighted average  …

     Notes from original source General comments  \
1013                        NaN              NaN

                                                 Source  \
1013  World Health Organization, Global Health Obser…

     Statistical concept and methodology  \
1013                                 NaN

                                  Development relevance Related source links  \
1013  Mortality rates due to unintentional poisoning…                  NaN

     Other web links  Related indicators License Type Unnamed: 20
1013             NaN                 NaN    CC BY-4.0         NaN

[1 rows x 21 columns]
```

[136]:
```
indicator2
```

[136]:
```
        Series Code              Topic  \
986  SH.STA.AIRP.P5  Health: Mortality

                                        Indicator Name Short definition  \
986  Mortality rate attributed to household and amb…              NaN

                                       Long definition Unit of measure  \
986  Mortality rate attributed to household and amb…             NaN
```

```
   Periodicity Base Period Other notes Aggregation method  …  \
986      Annual         NaN         NaN   Weighted average  …

   Notes from original source General comments  \
986                        NaN              NaN

                                             Source  \
986  World Health Organization, Global Health Obser…

   Statistical concept and methodology  \
986                                NaN

                               Development relevance Related source links  \
986  Air pollution is one of the biggest environmen…                  NaN

   Other web links  Related indicators License Type Unnamed: 20
986             NaN                NaN    CC BY-4.0         NaN

[1 rows x 21 columns]
```

[137]:

```python
# Define the indicator names you are interested in
indicator1 = "Mortality rate attributed to household and ambient air pollution,␣
↪age-standardized (per 100,000 population)"
indicator2 = "Mortality rate attributed to unintentional poisoning (per 100,000␣
↪population)"

# Check if data is available for the first indicator
data1 = WDIData[WDIData['Indicator Name'] == indicator1]
import pandas as pd

# Assuming your 'data1' DataFrame contains columns: 'Country Name', 'Indicator␣
↪Name', and individual years as columns (e.g., '1990', '1991', ...)
# You can select the columns you want to melt by specifying them in the␣
↪'id_vars' parameter

# Define the ID columns (columns to keep as they are)
id_vars = ['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code']

# Melt the DataFrame
melted_data1 = pd.melt(data1, id_vars=id_vars, var_name='Year',␣
↪value_name='Mortality Rate')

melted_data1
# Merge 'pm25_data' with 'data1' on a common column, such as 'Country Name' and␣
↪'Year'
```

```
merged_data1 = melted_data1.merge(pm25_data, on=['Country Name', 'Year'],␣
↪how='inner')
merged_data1
```

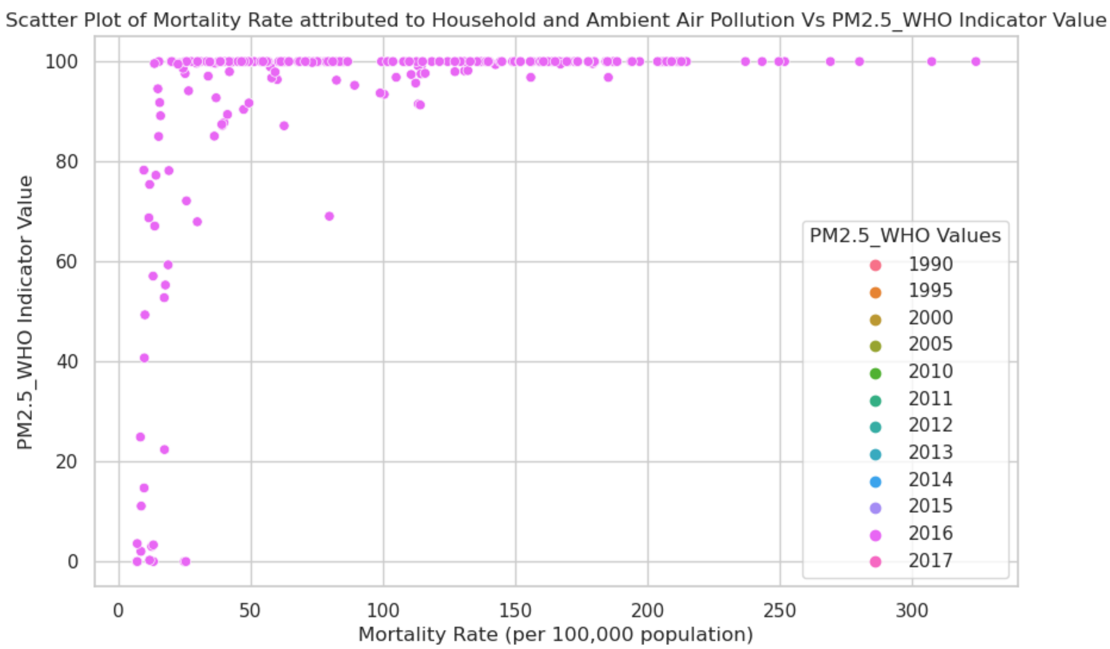
[137]:

```
                     Country Name Country Code_x  \
0                      Arab World            ARB
1          Caribbean small states            CSS
2      Central Europe and the Baltics        CEB
3        Early-demographic dividend          EAR
4             East Asia & Pacific            EAS
…                             …             …
2875         Virgin Islands (U.S.)           VIR
2876           West Bank and Gaza            PSE
2877                  Yemen, Rep.            YEM
2878                       Zambia            ZMB
2879                     Zimbabwe            ZWE

                                         Indicator Name_x Indicator Code_x  \
0     Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
1     Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
2     Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
3     Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
4     Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
…                                                  …              …
2875  Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
2876  Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
2877  Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
2878  Mortality rate attributed to household and amb…   SH.STA.AIRP.P5
2879  Mortality rate attributed to household and amb…   SH.STA.AIRP.P5

      Year  Mortality Rate Country Code_y Indicator Name_y   Indicator Code_y  \
0     1990             NaN            ARB        PM2.5_WHO  EN.ATM.PM25.MC.ZS
1     1990             NaN            CSS        PM2.5_WHO  EN.ATM.PM25.MC.ZS
2     1990             NaN            CEB        PM2.5_WHO  EN.ATM.PM25.MC.ZS
3     1990             NaN            EAR        PM2.5_WHO  EN.ATM.PM25.MC.ZS
4     1990             NaN            EAS        PM2.5_WHO  EN.ATM.PM25.MC.ZS
…      …               …             …              …                 …
2875  2017             NaN            VIR        PM2.5_WHO  EN.ATM.PM25.MC.ZS
2876  2017             NaN            PSE        PM2.5_WHO  EN.ATM.PM25.MC.ZS
2877  2017             NaN            YEM        PM2.5_WHO  EN.ATM.PM25.MC.ZS
2878  2017             NaN            ZMB        PM2.5_WHO  EN.ATM.PM25.MC.ZS
2879  2017             NaN            ZWE        PM2.5_WHO  EN.ATM.PM25.MC.ZS

      Indicator Value
0          100.000000
1          100.000000
2           98.945833
```

```
3            99.778256
4            98.381801
…                 …
2875         40.000000
2876        100.000000
2877        100.000000
2878        100.000000
2879        100.000000

[2880 rows x 10 columns]
```

[138]:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Create a scatter plot with 'Year' on the x-axis, 'Mortality Rate' on the␣
 ↪y-axis, and 'PM2.5_WHO' values as the hue
plt.figure(figsize=(10, 6))
sns.scatterplot(data=merged_data1, x='Mortality Rate', y='Indicator Value',␣
 ↪hue='Year')

# Label the axes
plt.xlabel('Mortality Rate (per 100,000 population)')
plt.ylabel('PM2.5_WHO Indicator Value')

# Set the title
plt.title('Scatter Plot of Mortality Rate attributed to Household and Ambient␣
 ↪Air Pollution Vs PM2.5_WHO Indicator Value')

# Show the grid
plt.grid(True)

# Show the plot
plt.legend(title='PM2.5_WHO Values')
plt.show()
```



10.3 (4 marks)

Extend the analysis above to find some countries of interest. These are defined as

The countries that have a high mortality rate due to household and ambient air pollution, but with low PM2.5 exposure

The countries that have a low mortality rate due to household and ambient air pollution, but with high PM2.5 exposure

Answer.

[94]:
```
#I  made these thresholds based on data that I've seen in this assignment
# Define your threshold values
high_mortality_threshold = 30
low_pm25_threshold = 75

low_mortality_threshold = 20
high_pm25_threshold = 40

# Find countries with high mortality rate and low PM2.5 exposure
high_mortality_low_pm25 = merged_data1[(merged_data1['Mortality Rate'] >␣
↪high_mortality_threshold) & (merged_data1['Indicator Value'] <␣
↪low_pm25_threshold)]

# Find countries with low mortality rate and high PM2.5 exposure
```

```
low_mortality_high_pm25 = merged_data1[(merged_data1['Mortality Rate'] <␣
↪low_mortality_threshold) & (merged_data1['Indicator Value'] >␣
↪high_pm25_threshold)]

# Display the results
print("Countries with high mortality rate and low PM2.5 exposure:")
print(high_mortality_low_pm25[['Country Name', 'Year', 'Mortality Rate',␣
↪'Indicator Value']])

print("\nCountries with low mortality rate and high PM2.5 exposure:")
print(low_mortality_high_pm25[['Country Name', 'Year', 'Mortality Rate',␣
↪'Indicator Value']])
```

```
Countries with high mortality rate and low PM2.5 exposure:
     Country Name  Year  Mortality Rate  Indicator Value
2607    Sri Lanka  2016            79.8        69.062963

Countries with low mortality rate and high PM2.5 exposure:
                   Country Name  Year  Mortality Rate  Indicator Value
2407                  Euro area  2016       14.252916        77.257491
2411             European Union  2016       19.163448        78.180420
2414                High income  2016       17.836323        55.300719
2433               OECD members  2016       18.875826        59.329538
2436  Post-demographic dividend  2016       17.401723        52.769713
2456                    Austria  2016       15.300000        85.015935
2458               Bahamas, The  2016       19.900000       100.000000
2463                    Belgium  2016       15.700000        91.789115
2494                    Denmark  2016       13.200000        57.091762
2508                     France  2016        9.700000        78.277498
2512                    Germany  2016       16.000000        89.154663
2531                     Israel  2016       15.400000       100.000000
2532                      Italy  2016       15.000000        94.548484
2534                      Japan  2016       11.900000        75.412111
2550                 Luxembourg  2016       11.600000        68.732730
2570                Netherlands  2016       13.700000        99.595798
2606                      Spain  2016        9.900000        40.724385
2613                Switzerland  2016       10.100000        49.339454
2628             United Kingdom  2016       13.800000        67.100673
```

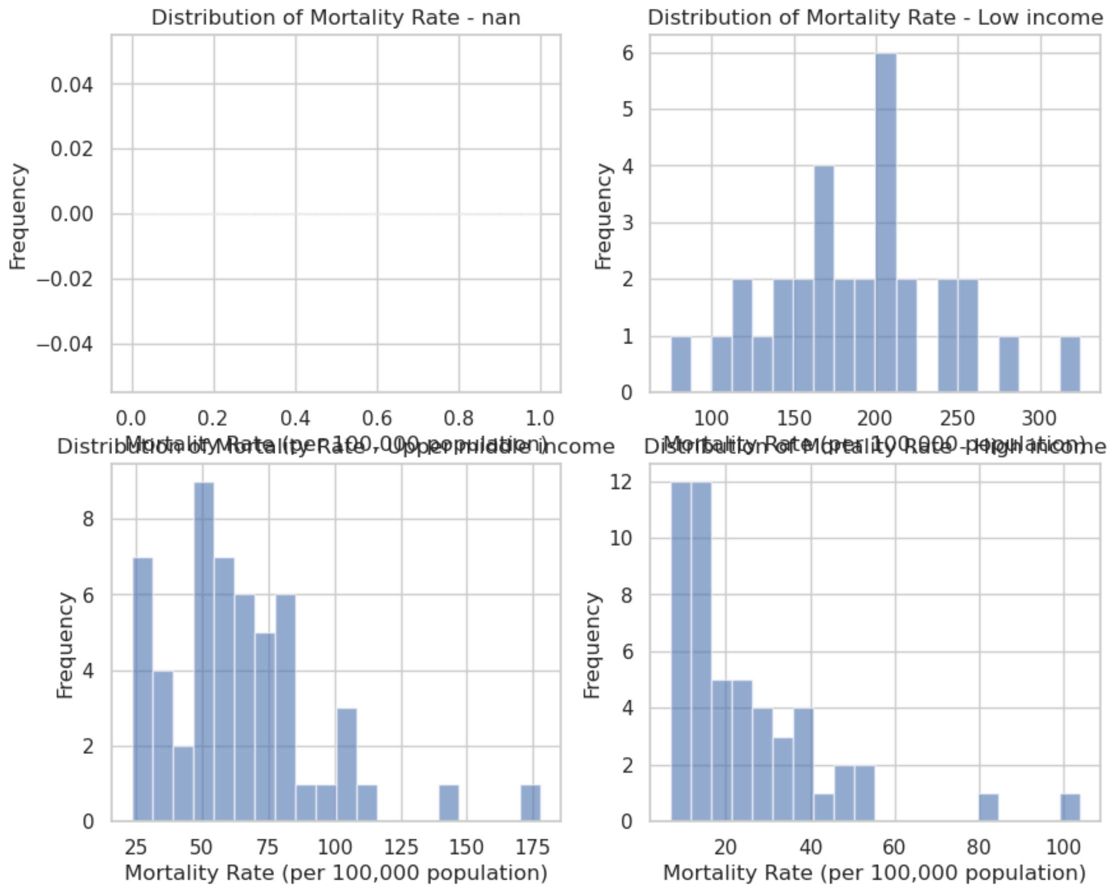
10.4 (10 marks)

Finally, we want to look at the mortality data by income. We expect higher income countries to have lower pollution-related mortality. Find out if this assumption holds. Calculate summary statistics and histograms for each income category and note any trends.

[133]:
```python
# Merge 'pm25_with_income' with 'merged_data1'
final_merged_data = merged_data1.merge(pm25_with_income, on=['Country Name',␣
↪'Year'], how='inner')

# Display the merged data
final_merged_data

income_groups = final_merged_data['Income Group'].unique()

plt.figure(figsize=(10, 8))

for i, income_group in enumerate(income_groups):
    plt.subplot(2, 2, i + 1)  # Create subplots
    group_data = final_merged_data[final_merged_data['Income Group'] ==␣
↪income_group]
    plt.hist(group_data['Mortality Rate'], bins=20, alpha=0.6)
    plt.xlabel('Mortality Rate (per 100,000 population)')
    plt.ylabel('Frequency')
    plt.title(f'Distribution of Mortality Rate - {income_group}')
    plt.grid(True)

plt.show()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[133], line 12
      9 plt.figure(figsize=(10, 8))
     11 for i, income_group in enumerate(income_groups):
---> 12     plt.subplot(2, 2, i + 1)  # Create subplots
     13     group_data = final_merged_data[final_merged_data['Income Group'] ==␣
↪income_group]
     14     plt.hist(group_data['Mortality Rate'], bins=20, alpha=0.6)

File /opt/conda/lib/python3.11/site-packages/matplotlib/pyplot.py:1323, in␣
↪subplot(*args, **kwargs)
   1320 fig = gcf()
   1322 # First, search for an existing subplot with a matching spec.
-> 1323 key = SubplotSpec._from_subplot_args(fig, args)
   1325 for ax in fig.axes:
   1326     # if we found an Axes at the position sort out if we can re-use it
   1327     if ax.get_subplotspec() == key:
   1328         # if the user passed no kwargs, re-use

File /opt/conda/lib/python3.11/site-packages/matplotlib/gridspec.py:600, in␣
↪SubplotSpec._from_subplot_args(figure, args)
    598 else:
    599     if not isinstance(num, Integral) or num < 1 or num > rows*cols:
```

```
--> 600         raise ValueError(
    601             f"num must be an integer with 1 <= num <= {rows*cols}, "
    602             f"not {num!r}"
    603         )
    604     i = j = num
    605 return gs[i-1:j]

ValueError: num must be an integer with 1 <= num <= 4, not 5
```



Answer.

10.5 (8 marks)

At the start, we asked some questions. Based on your analysis, provide a short answer to each of these:

Are we making any progress in reducing the amount of emitted pollutants across the globe?

Which are the critical regions where we should start environmental campaigns?

Are we making any progress in the prevention of deaths related to air pollution?

Which demographic characteristics seem to correlate with the number of health-related issues derived from air pollution?

Answer.

1) Are we making any progress in reducing the amount of emitted pollutants across the globe?

I don't think the top 5 countries are doing enough to reduce emissions. Even if other countries around the globe might be, these 5 make up 50%+ of global emissions, so they really have to reduce their emissions for global emissions to have some progress.

2) Which are the critical regions where we should start environmental campaigns?

I think that environmental campaigns should start in the top 5 countries for emissions since it is clear that they are making a big impact on gloabl emissions by contributing towards over 50% of gloabl emissions. Hopefully these environemntal campaigns can help the countries reduce emissions and improve overal global emissions.

3) Are we making any progress in the prevention of deaths related to air pollution?

I don't think there is much progress being made currently to prevent the deaths related to air pollution because of the mortality rates that we observed in this analysis. Countries must start to collectively reduce emissions and increase healthcare access to prevent deaths due to air pollution.

4) Which demographic characteristics seem to correlate with the number of health-related issues derived from air pollution?

Based on this analysis, we can see that income level is a key demographic characteristic that correlates with health-related issues from air pollution. This is because overall, lower income groups tend to experience higher mortality rates and more PM2.5 exposure. Also, lower income families lack the resources to get access to healthcare to avoid getting sick or hacing air pollution related health issues.
